Federated Incremental Semantic Segmentation

Jiahua Dong[1, 2, 3]*, Duzhen Zhang[4]*, Yang Cong[1, 2]†, Wei Cong[1, 2, 3], Henghui Ding[4], Dengxin Dai[4]
[1]State Key Laboratory of Robotics, Shenyang Institute of Automation,
Chinese Academy of Sciences, Shenyang, 110016, China.‡
[2]Institutes for Robotics and Intelligent Manufacturing,
Chinese Academy of Sciences, Shenyang, 110169, China.
[3]University of Chinese Academy of Sciences, Beijing, 100049, China.
[4]ETH Zürich, Zürich, 8092, Switzerland.
{dongjiahua1995, congyang81, congwei45, henghui.ding}@gmail.com, dai@vision.ee.ethz.ch

Abstract

Federated learning-based semantic segmentation (FSS) has drawn widespread attention via decentralized training on local clients. However, most FSS models assume categories are fixed in advance, thus heavily undergoing forgetting on old categories in practical applications where local clients receive new categories incrementally while have no memory storage to access old classes. Moreover, new clients collecting novel classes may join in the global training of FSS, which further exacerbates catastrophic forgetting. To surmount the above challenges, we propose a ***F****orgetting-* ***B****alanced* ***L****earning (****FBL****) model to address heterogeneous forgetting on old classes from both intra-client and inter-client aspects. Specifically, under the guidance of pseudo labels generated via adaptive class-balanced pseudo labeling, we develop a forgetting-balanced semantic compensation loss and a forgetting-balanced relation consistency loss to rectify intra-client heterogeneous forgetting of old categories with background shift. It performs balanced gradient propagation and relation consistency distillation within local clients. Moreover, to tackle heterogeneous forgetting from inter-client aspect, we propose a task transition monitor. It can identify new classes under privacy protection and store the latest old global model for relation distillation. Qualitative experiments reveal large improvement of our model against comparison methods. The code is available at https://github.com/JiahuaDong/FISS.*

1. Introduction

Federated learning (FL) [13,20,22,44] is a remarkable decentralized training paradigm to learn a global model across distributed local clients without accessing their private data. Under privacy preservation, it has achieved rapid development in semantic segmentation [4, 8, 30] by training on multiple decentralized local clients to alleviate the constraint of data island that requires enormous finely-labeled pixel annotations [25]. As a result, federated learning-based semantic segmentation (FSS) [28,29] significantly economizes annotation costs in data-scarce scenarios via training a global segmentation model on private data of different clients [29].

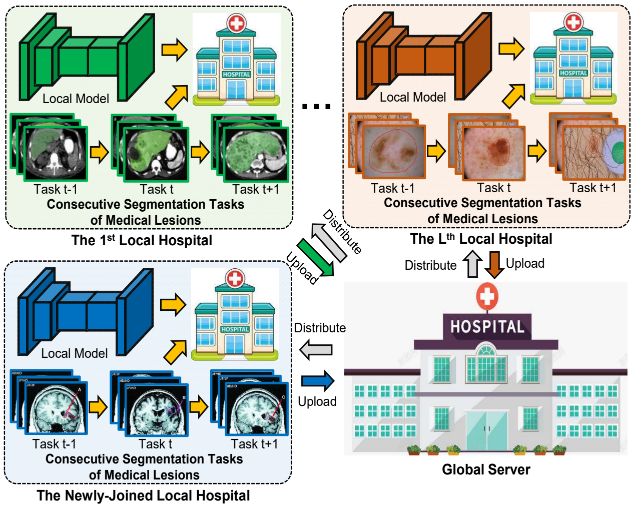

Figure 1. Exemplary FISS setting for medical diagnosis. Hundreds of hospitals including newly-joined ones receive new classes incrementally according to their own preference. FISS aims to segment new diseases consecutively via collaboratively learning a global segmentation model on private medical data of different hospitals.

However, existing FSS methods [16, 25, 28, 29] unrealistically assume that the learned foreground classes are static and fixed over time, which is impractical in real-world dynamic applications where local clients receive streaming data of new categories consecutively. To tackle this issue, existing

*Equal contributions. †The corresponding author is Prof. Yang Cong.

‡This work was supported in part by the National Nature Science Foundation of China under Grant 62127807, 62225310 and 62133005.

FSS methods [28, 29, 35] typically enforce local clients to store all samples of previously-learned old classes, and then learn a global model to segment new categories continually via FL. Nevertheless, it requires large computation and memory overhead as new classes arrive continuously, limiting the application ability of FSS methods [16, 28]. If local clients have no memory to store old classes, existing FSS methods [16, 29] significantly degrade segmentation behavior on old categories (*i.e.*, catastrophic forgetting [40, 47, 48]) when learning new classes incrementally. In addition, the pixels labeled as background in the current learning task may belong to old classes from old tasks or new foreground classes from future tasks. This phenomenon is also known as background shift [11, 36] that heavily aggravates heterogeneous forgetting speeds on old categories. More importantly, in practical scenarios, new local clients receiving new categories incrementally may join in global FL training irregularly, thus further exacerbating catastrophic forgetting to some extent.

To surmount the above real-world scenarios, we propose a novel practical problem called **F**ederated **I**ncremental **S**emantic **S**egmentation (**FISS**), where local clients collect new categories consecutively according to their preferences, and new local clients collecting unseen novel classes participate in global FL training irregularly. In the FISS settings, the class distributions are non-independent and identically distributed (Non-IID) across different clients, and training data of old classes is unavailable for all local clients. FISS aims to train a global incremental segmentation model via collaborative FL training on local clients while addressing catastrophic forgetting. In this paper, we use medical lesions segmentation [25, 29] as an example to better illustrate FISS, as shown in Figure 1. Hundreds of hospitals, as well as newly joined ones, collect unseen/new medical lesions continuously in clinical diagnosis. Considering privacy preservation, it is desired for these hospitals to learn a global segmentation modal via FL without accessing each other's data [44, 56].

A naive solution for FISS problem is to directly integrate incremental semantic segmentation [1, 11, 53] and FL [19, 50] together. Nevertheless, such a trivial solution requires global server to have strong human prior about which and when local clients can collect new categories, so that global model learned in the latest old task can be stored by local clients to address forgetting on old classes via knowledge distillation [18, 43]. Considering privacy preservation in the FISS, this privacy-sensitive prior knowledge cannot be shared between local clients and global server. As a result, this naive solution severely suffers from intra-client heterogeneous forgetting on different old classes caused by background shift [1, 11, 36, 53], and inter-client heterogeneous forgetting across different clients brought by Non-IID class distributions.

To overcome the above-mentioned challenges, we develop a novel **F**orgetting-**B**alanced **L**earning (**FBL**) model, which alleviates heterogeneous forgetting on old classes from intra-client and inter-client perspectives. Specifically, to tackle intra-client heterogeneous forgetting caused by background shift, we propose an adaptive class-balanced pseudo labeling to adaptively generate confident pseudo labels for old classes. Under the guidance of pseudo labels, we propose a forgetting-balanced semantic compensation loss to rectify different forgetting of old classes with background shift via considering balanced gradient propagation of local clients. In addition, a forgetting-balanced relation consistency loss is designed to distill underlying category-relation consistency between old and new classes for intra-client heterogeneous forgetting compensation. Moreover, considering addressing heterogeneous forgetting from inter-client aspect, we develop a task transition monitor to automatically identify new classes without any human prior, and store the latest old model from global perspective for relation consistency distillation. Experiments on segmentation datasets reveal large improvement of our model over comparison methods. We summarize the main contributions of this work as follows:

- We propose a novel practical problem called Federated Incremental Semantic Segmentation (FISS), where the major challenges are intra-client and inter-client heterogeneous forgetting on old categories caused by intra-client background shift and inter-client Non-IID distributions.
- We propose a Forgetting-Balanced Learning (FBL) model to address the FISS problem via surmounting heterogeneous forgetting from both intra-client and inter-client aspects. As we all know, in the FL field, this is a pioneer attempt to explore a global continual segmentation model.
- We develop a forgetting-balanced semantic compensation loss and a forgetting-balanced relation consistency loss to tackle intra-client heterogeneous forgetting across old classes, under the guidance of confident pseudo labels generated via adaptive class-balanced pseudo labeling.
- We design a task transition monitor to surmount inter-client heterogeneous forgetting by accurately recognizing new classes under privacy protection and storing the latest old model from global aspect for relation distillation.

2. Related Work

Federated Learning (FL) [26, 34, 44, 49] aggregates local-client model parameters to optimize a global model under privacy protection. [41] enforces local model to approximate the global ones via a proximal term. To minimize computation cost, [6] employs a layer-wise parameter aggregation strategy. Inspired by above FL [13, 44, 52] methods, [25, 28, 29] apply FL to semantic segmentation [5, 30], which has achieved rapid developments in medical analysis [9, 35] and autonomous driving [16]. [38] considers adversarial framework [57, 58] to tackle domain adaptation problem [15, 23, 24, 46] in the FL field. [10] proposes a federated class-incremental learning model via considering global and local forgetting. However, the above-mentioned methods [16, 29, 39] cannot segment new foreground classes continuously under the FISS settings.

Incremental Semantic Segmentation (ISS) [1, 11, 37, 53] considers class-incremental learning [21, 31, 42] in semantic segmentation. The key challenges of ISS are catastrophic forgetting [33, 40] and background shift [11, 37], as claimed in [1, 32]. ILT [36] proposes to distill latent features and probabilities between old and new models. PLOP [11] utilizes multi-scale pooling distillation to maintain past experience. SDR [37] considers feature consistency by prototype matching and contrastive learning [3]. RCIL [53] decouples the network into branches to overcome forgetting. Considering tackling background shift, [2, 11, 51] propose pseudo labeling to annotate old classes labeled as background pixels. Nevertheless, these ISS methods [11, 51, 53] cannot be effectively applied to address the FISS problem, due to their strong prior knowledge to access privately-sensitive information (*i.e.*, when and which local clients receive new classes).

3. Problem Definition

As claimed in incremental semantic segmentation (ISS) [1, 11, 27, 36], some consecutive segmentation tasks are defined as $\mathcal{T} = \{\mathcal{T}^t\}_{t=1}^T$, where the t-th ($t = 1, \cdots, T$) task $\mathcal{T}^t = \{\mathbf{x}_i^t, \mathbf{y}_i^t\}_{i=1}^{N^t}$ is composed of N^t pairs of RGB images $\mathbf{x}_i^t \in \mathbb{R}^{H \times W \times 3}$ and labels $\mathbf{y}_i^t \in \mathbb{R}^{H \times W}$. H and W denote height and width of given images. The label space $\mathcal{Y}^t$ of t-th incremental task consists of K^t new categories and background. K^t new classes have no overlap with $K^o = \sum_{i=1}^{t-1} K^i \subset \cup_{j=1}^{t-1} \mathcal{Y}^j$ old classes learned from $t-1$ old tasks. In the t-th task, we follow ISS methods [11, 36] to annotate K^o old classes and other foreground classes from future learning tasks as background (*i.e.*, background shift [1]), due to unavailable training data of K^o old classes.

We then extend the settings of incremental semantic segmentation (ISS) [1, 36, 53] to Federated Incremental Semantic Segmentation (FISS). Denote global server as $\mathcal{S}_g$ and L local clients as $\{\mathcal{S}_l\}_{l=1}^L$. In the FISS, at the r-th ($r = 1, \cdots, R$) global round, we randomly select some local clients to aggregate gradients. When we choose the l-th local client to learn the t-th segmentation task, the latest global model $\Theta^{r,t}$ is distributed to $\mathcal{S}_l$, and trained on private training data $\mathcal{T}_l^t = \{\mathbf{x}_{li}^t, \mathbf{y}_{li}^t\}_{i=1}^{N_l^t} \sim \mathcal{P}_l$ of $\mathcal{S}_l$. $\mathbf{x}_{li}^t$ and $\mathbf{y}_{li}^t \in \mathcal{Y}_l^t$ denote the images and labels of the l-th client. $\{\mathcal{P}_l\}_{l=1}^L$ are non-independent and identically distributed (*i.e.*, Non-IID) across local clients. The label space $\mathcal{Y}_l^t \subset \mathcal{Y}^t$ of $\mathcal{S}_l$ in the t-th task is composed of K_l^t new classes ($K_l^t \leq K^t$) that belongs to a subset of $\mathcal{Y}^t = \cup_{l=1}^L \mathcal{Y}_l^t$. Following ISS methods [11, 36, 53], we consider background shift in the FISS and also annotate $K_l^o = \sum_{i=1}^{t-1} K_l^i \subset \cup_{j=1}^{t-1} \mathcal{Y}_l^j$ old categories from $t-1$ old tasks and other foreground categories from future learning tasks as background. After getting global model $\Theta^{r,t}$ and performing local training on $\mathcal{T}_l^t$, $\mathcal{S}_l$ obtains a updated local model $\Theta_l^{r,t}$. Then global server $\mathcal{S}_g$ aggregates local models of selected clients as the global model $\Theta^{r+1,t}$ for the training of next global round.

In the t-th task, motivated by [10], all local clients $\{\mathcal{S}_l\}_{l=1}^L$ are divided into three categories: $\{\mathcal{S}_l\}_{l=1}^L = \mathbf{S}_o \cup \mathbf{S}_c \cup \mathbf{S}_n$. Specifically, $\mathbf{S}_o$ is composed of L_o local clients that have accumulated past experience for previous tasks but cannot collect new training data of the t-th task; $\mathbf{S}_c$ consisting of L_c local clients can receive new training data of current task and has learning experience for old classes; $\mathbf{S}_n$ includes L_n new local clients with unseen novel classes but without past learning experience of old classes. These local clients are randomly determined in each incremental task. New clients $\mathbf{S}_n$ are added randomly at any global round in the FISS, increasing $L = L_o + L_c + L_n$ gradually as continuous tasks. More importantly, we don't have prior knowledge about the class distributions $\{\mathcal{P}_l\}_{l=1}^L$, quantity and order of segmentation tasks, when and which local clients receive new classes. In this paper, FISS aims to learn a global model $\Theta^{R,T}$ to segment new categories continuously while surmounting heterogeneous forgetting on old categories brought by background shift, under the privacy preservation of local clients.

4. The Proposed Model

Figure 2 presents the overview of our model to address the FISS problem. Our FBL model overcomes intra-client heterogeneous forgetting via a forgetting-balanced semantic compensation loss (Section 4.2) and a forgetting-balanced relation consistency loss (Section 4.3), under the guidance of adaptive class-balanced pseudo labeling (Section 4.1) to mine pseudo labels for old classes with background shift. Meanwhile, it addresses inter-client heterogeneous forgetting via a task transition monitor (Section 4.4) to recognize new classes and store old model for relation distillation.

4.1. Adaptive Class-Balanced Pseudo Labeling

For the l-th local client $\mathcal{S}_l \in \mathbf{S}_c \cup \mathbf{S}_n$, the semantic segmentation loss $\mathcal{L}_{\mathrm{SE}}$ for a mini-batch $\{\mathbf{x}_{li}^t, \mathbf{y}_{li}^t\}_{i=1}^B \subset \mathcal{T}_l^t$ sampled from the t-th incremental task is formulated as:

$$\mathcal{L}_{\mathrm{SE}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{HW} \mathcal{D}_{\mathrm{CE}}(\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta^{r,t})_j, (\mathbf{y}_{li}^t)_j), \tag{1}$$

where $\mathcal{D}_{\mathrm{CE}}(\cdot, \cdot)$ denotes the cross-entropy loss. At the r-th global round, global model $\Theta^{r,t}$ is transmitted from global server $\mathcal{S}_g$ to $\mathcal{S}_l$. $\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta^{r,t})_j \in \mathbb{R}^{1+K^o+K^t}$ is the probability at the j-th ($j = 1, \cdots, HW$) pixel predicted by $\Theta^{r,t}$, and it predicts background, K^o old classes, and K^t new classes for the j-th pixel. $(\mathbf{y}_{li}^t)_j \in \mathcal{Y}_l^t$ is corresponding label at the j-th pixel. B represents the batch size.

As aforementioned, in the FISS settings, local client $\mathcal{S}_l$ has no memory to store K^o old classes, while back-

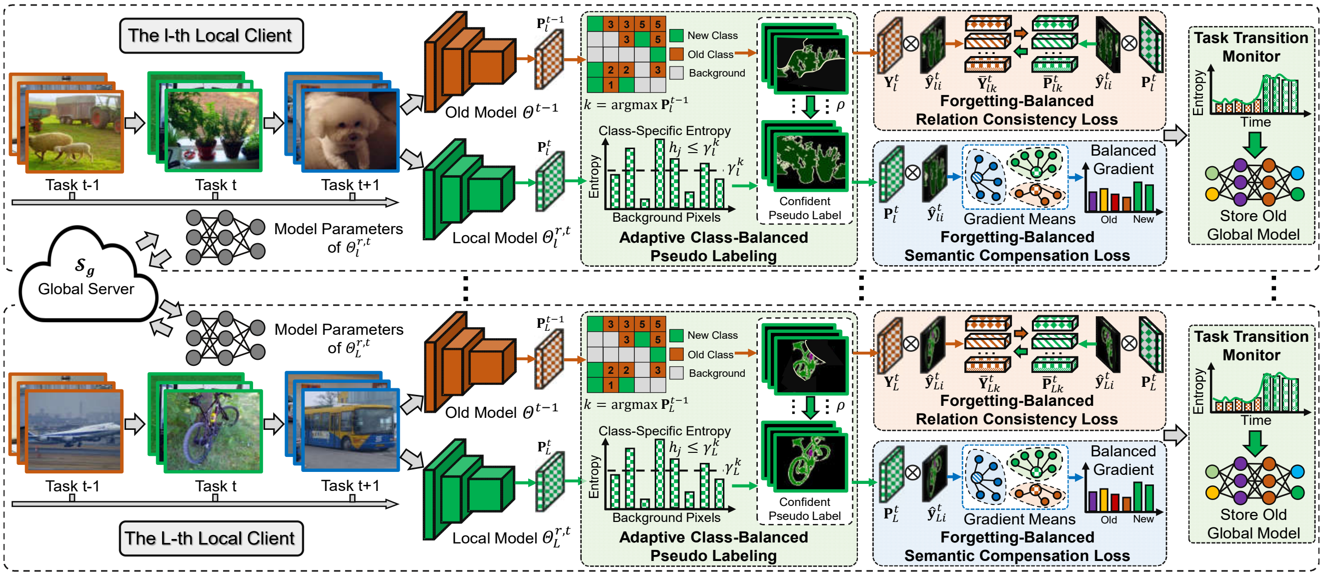

Figure 2. Overview of the proposed FBL model. It includes a *forgetting-balanced semantic compensation loss* $\mathcal{L}_{\mathrm{FS}}$ and a *forgetting-balanced relation consistency loss* $\mathcal{L}_{\mathrm{FR}}$ to tackle intra-client heterogeneous forgetting brought by background shift, under the guidance of *adaptive class-balanced pseudo labeling*. Meanwhile, it utilizes a *task transition monitor* to overcome inter-client heterogeneous forgetting brought by Non-IID distributions with background shift.

ground pixels may belong to K^o old classes, other foreground classes from future tasks or real background (*i.e.*, background shift [1, 11]). As a result, it enforces the updating of local model $\Theta_l^{r,t}$ (*i.e.*, Eq. (1)) to suffer from intra-client heterogeneous forgetting among different old classes brought by background shift, after $\mathcal{S}_l$ receives the global model $\Theta^{r,t}$ from $\mathcal{S}_g$ for local training. To this end, as shown in Figure 2, we develop an adaptive class-balanced pseudo labeling to adaptively mine confident pseudo labels for old classes labeled as background pixels in the t-th segmentation task. Different from existing ISS methods [2, 11, 51] that only utilize a constant probability threshold to select pseudo labels for all classes, our FBL model considers class balance to mine pseudo labels for old classes via introducing class-specific entropy threshold for each old class, which are adaptively determined as continual learning process. These class-balanced pseudo labels of K^o old classes are essential to alleviate heterogeneous forgetting across different old classes within local clients.

In the t-th task, as shown in Figure 2, given a sample $\{\mathbf{x}_{li}^t, \mathbf{y}_{li}^t\} \subset \mathcal{T}_l^t$, we feed it into old global model Θ^{t-1} of the last task and current local model $\Theta_l^{r,t}$ to obtain the probabilities $\mathbf{P}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1}) \in \mathbb{R}^{H\times W\times(1+K^o)}$ and $\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta_l^{r,t}) \in \mathbb{R}^{H\times W\times(1+K^o+K^t)}$ respectively. Then pseudo label $\hat{\mathbf{y}}_{li}^t \in \mathbb{R}^{H\times W}$ of given image $\mathbf{x}_{li}^t$ is defined as:

$$(\hat{\mathbf{y}}_{li}^t)_j = \begin{cases} k, \text{ if } (\mathbf{y}_{li}^t)_j \notin \mathcal{Y}_l^b \text{ and } k = (\mathbf{y}_{li}^t)_j; \\ k, \text{ if } (\mathbf{y}_{li}^t)_j \in \mathcal{Y}_l^b \text{ and } h_j \leq \gamma_l^k \\ \quad \text{and } k = \arg\max \mathbf{P}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1})_j; \\ 0, \text{ otherwise}, \end{cases} \tag{2}$$

Algorithm 1: Determination of $\{\gamma_l^k\}_{k=1}^{K^o}$ in Eq. (2).

```
Input: $\mathcal{T}_l^t = \{\mathbf{x}_{li}^t, \mathbf{y}_{li}^t\}_{i=1}^{N_l^t}$, and the selection proportion $\rho$;
for $i = 1, \cdots, N_l^t$ do
    $\mathbf{H}_{li}^t = \mathcal{H}(\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta_l^{r,t})) \in \mathbb{R}^{H\times W}$;
    $\mathbf{L}_{li}^t = \arg\max \mathbf{P}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1}) \in \mathbb{R}^{H\times W}$;
    for $k = 1, \cdots, K^o$ do
        $\mathbf{H}_l^k = \mathbf{H}_{li}^t[\mathbf{L}_{li}^t == k]$;
        $\mathbf{M}_l^k = [\mathbf{M}_l^k; \mathrm{matrix\_to\_vector}(\mathbf{H}_l^k)]$;
for $k = 1, \cdots, K^o$ do
    $\mathbf{E}_l^k = \mathrm{sort}(\mathbf{M}_l^k, \mathrm{order} = \mathrm{ascending})$;
    $\gamma_l^k = \mathbf{E}_l^k[\mathrm{length}(\mathbf{E}_l^k) \cdot \rho]$.
```

where $(\hat{\mathbf{y}}_{li}^t)_j$ is pseudo label of the j-th pixel from $\hat{\mathbf{y}}_{li}^t$. $\mathbf{P}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1})_j$ is softmax probability of the j-th pixel from $\mathbf{P}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1})$. $h_j = \mathcal{H}(\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta_l^{r,t})_j)$ represents entropy of the j-th pixel, and $\mathcal{H}(\mathbf{p}) = \sum_i \mathbf{p}_i \log \mathbf{p}_i$ is entropy measure function. $\{\gamma_l^k\}_{k=1}^{K^o}$ denote class-specific entropy threshold to adaptively select class-balanced pseudo labels with high confidence. As shown in Eq. (2), in the t-th task $\mathcal{T}_l^t$, when the j-th pixel belongs to background label space $\mathcal{Y}_l^b$ (*i.e.*, $(\mathbf{y}_{li}^t)_j \in \mathcal{Y}_l^b$) and the entropy h_j is less than γ_l^k, pseudo label is determined by $(\hat{\mathbf{y}}_{li}^t)_j = \arg\max \mathbf{P}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1})_j$. If the j-th pixel is not labeled as background (*i.e.*, $(\mathbf{y}_{li}^t)_j \notin \mathcal{Y}_l^b$), we consider pseudo label as new foreground classes: $(\hat{\mathbf{y}}_{li}^t)_j = (\mathbf{y}_{li}^t)_j$. Otherwise, $(\hat{\mathbf{y}}_{li}^t)_j = 0$ denotes real background for the j-th pixel of $\hat{\mathbf{y}}_{li}^t$.

The determination of $\{\gamma_l^k\}_{k=1}^{K^o}$ is summarized in **Algorithm 1**. After computing entropy $\{\mathbf{H}_{li}^t\}_{i=1}^{N_l^t}$ for all samples in the t-th task $\mathcal{T}_l^t$, we sort the entropy of all pixels predicted as the k-th class. γ_l^k is determined via the entropy ranked at $[\text{length}(\mathbf{E}_l^k) \cdot \rho]$ of $\mathbf{E}_l^k$, where ρ is selection proportion for all old classes. The value of ρ is initialized as 20%, and adds 2% for each epoch empirically as training process. We set the maximum selection proportion ρ as 50%. Given a mini-batch $\{\mathbf{x}_{li}^t, \mathbf{y}_{li}^t\}_{i=1}^B \subset \mathcal{T}_l^t$, we generate class-balanced pseudo labels $\{\mathbf{x}_{li}^t, \hat{\mathbf{y}}_{li}^t\}_{i=1}^B \subset \mathcal{T}_l^t$ adaptively via considering class-balanced selection proportion ρ in Eq. (2) for all old classes. These confident pseudo labels provide strong guidance for the forgetting-balanced semantic compensation loss (Section 4.2) and forgetting-balanced relation consistency loss (Section 4.3) to surmount intra-client heterogeneous forgetting among different old classes.

4.2. Forgetting-Balanced Semantic Compensation

To address heterogeneous forgetting speeds of different old classes within local client $\mathcal{S}_l \in \mathbf{S}_c \cup \mathbf{S}_n$, we propose a forgetting-balanced semantic compensation loss $\mathcal{L}_{\text{FS}}$, as shown in Figure 2. It considers balanced gradient propagation between different old tasks for intra-client heterogeneous forgetting compensation. Specifically, the loss $\mathcal{L}_{\text{FS}}$ employs gradient propagation means of different old tasks to measure the forgetting heterogeneity of old classes, and then reweights segmentation loss $\mathcal{L}_{\text{SE}}$ in Eq. (1) to normalize heterogeneous forgetting speeds brought by background shift. For a given sample $\{\mathbf{x}_{li}^t, \hat{\mathbf{y}}_{li}^t\} \subset \mathcal{T}_l^t$ with generated pseudo label, we first obtain its probability $\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta_l^{r,t})$ predicted via local model $\Theta_l^{r,t}$. Motivated by [45], we then formulate gradient scalar Γ_{ij}^t of the j-th pixel with respect to the k-th output neuron $\mathcal{N}_k^t$ of pixel classifier in $\Theta_l^{r,t}$ as follows:

$$\Gamma_{ij}^t = \frac{\partial \mathcal{D}_{\text{CE}}(\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta_l^{r,t})_j, (\hat{\mathbf{y}}_{li}^t)_j)}{\partial \mathcal{N}_k^t} = \mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta_l^{r,t})_j^k - 1, \quad (3)$$

where $\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta_l^{r,t})_j^k$ is probability of the k-th class at the j-th pixel of $\mathbf{x}_{li}^t$, and $k = (\hat{\mathbf{y}}_{li}^t)_j$ denotes pseudo label of the j-th pixel in $\mathbf{x}_{li}^t$. Considering that intra-client heterogeneous forgetting of old classes changes dynamically as continual learning tasks, we expect gradient scalar Γ_{ij}^t of old classes to be adaptive in the FISS, and reformulate Eq. (3) as:

$$\bar{\Gamma}_{ij}^t = |\Gamma_{ij}^t|^{\frac{K_l^o}{K_l^o + K_l^t}} \cdot \mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j \in \cup_{\eta=1}^{t-1} \mathcal{Y}_l^\eta} + |\Gamma_{ij}^t| \cdot \mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j \in \mathcal{Y}_l^t \cup \mathcal{Y}_l^b}. \quad (4)$$

where $\mathcal{Y}_l^b$ is background label space of the l-th local client $\mathcal{S}_l$. When pseudo label $(\hat{\mathbf{y}}_{li}^t)_j$ of the j-th pixel in $\mathbf{x}_{li}^t$ belongs to old classes from previous t–1 tasks, $\bar{\Gamma}_{ij}^t = |\Gamma_{ij}^t|^{K_l^o/(K_l^o + K_l^t)}$; otherwise, $\bar{\Gamma}_{ij}^t = |\Gamma_{ij}^t|$ for new classes and background.

As a result, given mini-batch samples $\{\mathbf{x}_{li}^t, \hat{\mathbf{y}}_{li}^t\}_{i=1}^B \in \mathcal{T}_l^t$ in the t-th segmentation task, we denote gradient propagation means Γ_b and Γ_η for the background and foreground classes learned from the η-th $(1 \leq \eta \leq t)$ task as follows:

$$\Gamma_b = \frac{1}{Z_b} \sum_{i=1}^{B} \sum_{j=1}^{HW} \bar{\Gamma}_{ij}^t, \quad \Gamma_\eta = \frac{1}{Z_\eta} \sum_{i=1}^{B} \sum_{j=1}^{HW} \bar{\Gamma}_{ij}^t, \quad (5)$$

where the quantity of pixels belonging to background and the η-th task are denoted as $Z_b = \sum_{i=1}^B \sum_{j=1}^{HW} \mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j \in \mathcal{Y}_l^b}$ and $Z_\eta = \sum_{i=1}^B \sum_{j=1}^{HW} \mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j \in \mathcal{Y}_l^\eta}$. The gradient propagation means Γ_b and $\{\Gamma_\eta\}_{\eta=1}^t$ in Eq. (5) reflect gradient-imbalanced propagation between old and new classes. Thus, these gradient means can effectively measure intra-client forgetting heterogeneity of different old classes, and evaluate updating speeds of new classes and background to some extent. Under the guidance of pseudo labels $\hat{\mathbf{y}}_{li}^t$, we employ $\{\Gamma_\eta\}_{\eta=1}^t$ and Γ_b to reweight semantic segmentation loss $\mathcal{L}_{\text{SE}}$, and formulate the proposed forgetting-balanced semantic compensation loss $\mathcal{L}_{\text{FS}}$ as follows:

$$\mathcal{L}_{\text{FS}} = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{HW} \frac{\bar{\Gamma}_{ij}^t}{\bar{\Gamma}} \cdot \mathcal{D}_{\text{CE}}\big(\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta_l^{r,t})_j, (\hat{\mathbf{y}}_{li}^t)_j\big), \quad (6)$$

where $\bar{\Gamma} = \sum_{\eta=1}^t \Gamma_\eta \cdot \mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j \in \mathcal{Y}_l^\eta} + \Gamma_b \cdot \mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j \in \mathcal{Y}_l^b}$ denotes different normalization weights for background, old and new classes. $\mathcal{L}_{\text{FS}}$ can address intra-client heterogeneous forgetting of different old classes via reweighting segmentation loss $\mathcal{L}_{\text{SE}}$ to achieve class-balanced gradient propagation.

4.3. Forgetting-Balanced Relation Consistency

The intrinsic relations between old and new classes are immutable in purely semantic space, independent of background shift [1, 11] and availability of training data of old classes. In light of this, consistent semantic relations between old model Θ^{t-1} and current local model $\Theta_l^{r,t}$ plays an important role in tackling intra-client heterogeneous forgetting on old classes. However, most existing ISS methods [1, 27, 36] only consider underlying relationships among old classes via performing knowledge distillation [18] on an individual sample, which can be severely affected by noisy predictions on old classes brought by background shift. In addition, forgetting heterogeneity of old classes within local clients enforces most ISS methods [1, 11] to suffer from heterogeneous inter-class relations distillation, thus aggravating imbalanced gradient propagation across incremental tasks.

To this end, we propose a forgetting-balanced relation consistency loss $\mathcal{L}_{\text{FR}}$ to tackle intra-client heterogeneous forgetting via compensating heterogeneous relation distillation. Specifically, we propose relationship prototype of each class instead of an individual sample to better characterize underlying relations between old and new classes, and consider gradient means in Eq. (5) to balance heterogeneous relation distillation. As shown in Figure 2, given $\{\mathbf{x}_{li}^t, \hat{\mathbf{y}}_{li}^t\}_{i=1}^B \subset \mathcal{T}_l^t$, we feed it into old model Θ^{t-1} and local model $\Theta_l^{r,t}$ of $\mathcal{S}_l$ to obtain probabilities $\mathbf{P}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1}) \in \mathbb{R}^{H \times W \times (1+K^o)}$ and $\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta_l^{r,t}) \in \mathbb{R}^{H \times W \times (1+K^o+K^t)}$. Then we substitute the first $1 + K^o$ channel dimensions of one-hot pseudo label $\hat{\mathbf{Y}}_{li}^t \in \mathbb{R}^{H \times W \times (1+K^o+K^t)}$ ($\hat{\mathbf{Y}}_{li}^t$ is one-hot encoding of $\hat{\mathbf{y}}_{li}^t$) with $\mathbf{P}_l^{t-1}(\mathbf{x}_{li}^t, \Theta^{t-1})$, and abbreviate this variant as

Table 1. Comparisons of mIoU (%) on Pascal-VOC 2012 dataset [12] under the setting of 15-1 with overlapped foregrounds.

Class ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	mIoU	Imp.
Finetuning + FL	70.7	6.5	0.0	0.0	11.2	0.1	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	9.7	4.7	⇑ 51.9
LWF [27] + FL	81.6	0.2	0.0	0.0	8.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.9	0.0	0.0	0.0	7.2	8.6	5.2	⇑ 51.4
ILT [36] + FL	82.3	13.1	0.0	0.0	8.2	0.0	5.5	0.0	0.0	3.2	0.3	11.3	0.0	17.6	0.1	1.8	0.0	0.0	5.7	7.8	13.0	8.1	⇑ 48.5
MiB [1] + FL	**84.9**	15.9	31.7	35.8	17.9	37.8	9.1	47.2	62.9	10.6	42.2	25.5	54.7	48.3	50.8	**77.7**	0.0	6.2	8.1	15.8	13.2	33.1	⇑ 23.5
PLOP [11] + FL	62.7	55.1	20.0	**49.6**	44.3	**60.1**	**82.4**	61.4	74.5	24.2	**43.7**	43.9	57.6	48.3	61.2	67.3	**14.6**	**44.4**	**10.4**	22.9	8.0	45.5	⇑ 11.1
RCIL [53] + FL	0.0	**76.0**	**41.9**	49.2	**63.4**	56.9	**84.2**	**82.5**	**85.3**	**36.5**	17.0	**55.7**	**74.6**	**64.2**	**78.8**	68.2	0.9	**29.0**	**15.3**	**43.0**	**28.3**	**50.0**	⇑ 6.6
FBL (Ours)	**88.7**	**81.9**	**37.3**	**79.1**	**60.5**	**71.3**	81.9	**79.7**	**81.9**	**34.6**	**58.3**	**57.0**	**70.3**	**70.4**	**79.4**	**80.5**	**1.8**	9.0	1.5	**40.5**	**23.6**	**56.6**	–

Table 2. Comparisons of mIoU (%) on Pascal-VOC 2012 dataset [12] under the setting of 4-4 with overlapped foregrounds.

Class ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	mIoU	Imp.
Finetuning + FL	73.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.0	20.1	**34.4**	**32.3**	30.4	9.1	⇑ 34.8
LWF [27] + FL	**88.4**	0.0	0.0	0.0	0.9	0.0	0.0	8.7	16.3	**8.5**	0.0	**39.2**	**39.6**	**38.6**	63.2	**77.7**	**24.9**	15.1	24.9	25.1	29.6	23.8	⇑ 20.1
ILT [36] + FL	**87.8**	0.0	0.0	0.0	8.5	0.1	4.1	22.7	14.5	2.4	0.0	**34.5**	25.2	36.0	63.5	**74.4**	**15.2**	13.5	23.4	24.7	26.0	22.7	⇑ 21.2
MiB [1] + FL	86.7	50.9	23.0	17.7	25.0	**8.9**	41.3	**67.1**	**47.9**	**4.5**	0.1	29.1	26.9	21.7	**69.8**	73.2	3.1	17.9	**30.3**	29.2	19.8	**33.0**	⇑ 10.9
PLOP [11] + FL	85.5	1.7	0.3	0.0	**44.3**	0.2	**66.1**	58.1	0.6	0.0	1.9	25.1	33.4	31.0	46.1	70.3	0.0	**27.5**	25.0	**36.1**	36.5	28.1	⇑ 15.8
RCIL [53] + FL	85.6	**62.8**	**29.6**	**38.9**	**39.3**	0.9	62.3	51.2	32.6	0.3	**34.1**	21.1	3.9	18.1	40.8	68.6	1.2	6.5	27.7	15.0	**39.1**	32.4	⇑ 11.5
FBL (Ours)	86.3	**66.2**	**34.0**	**48.3**	28.0	**6.9**	**64.7**	**75.6**	**74.1**	0.0	**26.0**	29.9	**61.7**	**40.1**	**66.0**	70.4	0.0	**40.4**	27.4	26.8	**48.5**	**43.9**	–

relationship label $\mathbf{Y}_l^t(\mathbf{x}_{li}^t, \Theta^{t-1}) \in \mathbb{R}^{H\times W\times(1+K^o+K^t)}$ indicating underlying relations among old and new categories. For the k-th class, the relationship prototype $\bar{\mathbf{P}}_{lk}^t$ and its label $\bar{\mathbf{Y}}_{lk}^t$ are written as follows:

$$\bar{\mathbf{P}}_{lk}^t = \frac{1}{Z_k}\sum_{i=1}^{B}\sum_{j=1}^{HW}\mathbf{P}_l^t(\mathbf{x}_{li}^t, \Theta_l^{r,t})_j \cdot \mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j=k}, \tag{7}$$

$$\bar{\mathbf{Y}}_{lk}^t = \frac{1}{Z_k}\sum_{i=1}^{B}\sum_{j=1}^{HW}\mathbf{Y}_l^t(\mathbf{x}_{li}^t, \Theta^{t-1}) \cdot \mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j=k}, \tag{8}$$

where $Z_k = \sum_{i=1}^{B}\sum_{j=1}^{HW}\mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j=k}$ is pixel number of the k-th class. The class-wise gradient mean Γ_k for the k-th class is formulated as $\Gamma_k = \frac{1}{Z_k}\sum_{i=1}^{B}\sum_{j=1}^{HW}\bar{\Gamma}_{ij}^t \cdot \mathbb{I}_{(\hat{\mathbf{y}}_{li}^t)_j=k}$, which is then used to reweight heterogeneous distillation gains. As a result, the forgetting-balanced relation consistency loss $\mathcal{L}_{\mathrm{FR}}$ is concretely written as follows:

$$\mathcal{L}_{\mathrm{FR}} = \frac{1}{K^o+K^t}\sum_{k=1}^{K^o+K^t}\frac{\Gamma_k}{\bar{\Gamma}_{\mathrm{cls}}} \cdot \mathcal{D}_{\mathrm{KL}}(\bar{\mathbf{P}}_{lk}^t, \bar{\mathbf{Y}}_{lk}^t), \tag{9}$$

where $\mathcal{D}_{\mathrm{KL}}(\cdot||\cdot)$ is Kullback-Leibler divergence. $\bar{\Gamma}_{\mathrm{cls}} = \sum_{\eta=1}^{t}\Gamma_\eta \cdot \mathbb{I}_{k\in\mathcal{Y}_l^\eta}$ denotes gradient normalization mean.

In summary, the major objective of the l-th local client $\mathcal{S}_l$ to learn the t-th segmentation task $\mathcal{T}_l^t$ is expressed as:

$$\mathcal{L}_{\mathrm{obj}} = \mathcal{L}_{\mathrm{FS}} + \lambda_1\mathcal{L}_{\mathrm{FR}} + \lambda_2\mathcal{L}_{\mathrm{POD}}, \tag{10}$$

where λ_1, λ_2 are trade-off parameters, and $\mathcal{L}_{\mathrm{POD}}$ denotes the local POD loss proposed in PLOP [11] to perform feature distillation. When $t \geq 2$, we set $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$ in Eq. (10) to train local model $\Theta_l^{r,t}$; otherwise, we utilize $\mathcal{L}_{\mathrm{SE}}$ in Eq. (1) to optimize $\Theta_l^{r,t}$. To learn the t-th segmentation task $\mathcal{T}_l^t$, local clients belonging to $\mathbf{S}_c$ and $\mathbf{S}_n$ share the same objective function (*i.e.*, Eq. (10)).

4.4. Task Transition Monitor

When local clients segment new classes consecutively via Eq. (10), global sever $\mathcal{S}_g$ requires to automatically identify when and which local clients collect new classes, and then store the latest old global model Θ^{t-1} to perform $\mathcal{L}_{\mathrm{FR}}$ and $\mathcal{L}_{\mathrm{POD}}$. As a result, the accurate selection of the latest old model Θ^{t-1} is essential to address inter-client heterogeneous forgetting across different local clients brought by Non-IID class distributions, when new foreground classes arrive. However, considering privacy preservation [54, 55], we don't have human prior about when to obtain new classes in local clients under the FISS settings. To address this challenge, a naive method is to detect whether the labels of current training data have been observed before. Nevertheless, the Non-IID distributions across local clients make it impossible to identify whether the collected data belongs to old classes seen by other clients or new categories. Thus, inspired by [10, 14], we design a task transition monitor to automatically recognize when and which local clients collect new categories. At the r-th round, when $\mathcal{S}_l$ receives global model $\Theta^{r,t}$, it evaluates the average entropy $\mathcal{I}_l^{r,t}$ on $\mathcal{T}_l^t$:

$$\mathcal{I}_l^{r,t} = \frac{1}{N_l^t}\sum_{i=1}^{N_l^t}\sum_{j=1}^{HW}\mathcal{H}(P_l^t(\mathbf{x}_{li}^t, \Theta^{r,t})_j), \tag{11}$$

where $\mathcal{H}(P_l^t(\mathbf{x}_{li}^t, \Theta^{r,t})) \in \mathbb{R}^{H\times W}$ is the entropy map of $\mathbf{x}_{li}^t$, and $\mathcal{H}(P_l^t(\mathbf{x}_{li}^t, \Theta^{r,t})_j)$ is entropy scalar of the j-th pixel. $\mathcal{H}(\mathbf{p}) = \sum_i \mathbf{p}_i \log \mathbf{p}_i$ is entropy measure function. We consider local clients are collecting new classes, if there is a sudden rise for averaged entropy $\mathcal{I}_l^{r,t}$: $\mathcal{I}_l^{r,t} - \mathcal{I}_l^{r-1,t} \geq \tau$. We then update t via $t \leftarrow t+1$, and automatically store the latest global model $\Theta^{r-1,t}$ at the $(r-1)$-th global round as old model Θ^{t-1} to optimize local model $\Theta_l^{r,t}$ via $\mathcal{L}_{\mathrm{obj}}$ in Eq. (10). We set $\tau = 0.6$ empirically in this paper. The automatic selection of old model Θ^{t-1} from global aspect is

Table 3. Comparisons of mIoU (%) on Pascal-VOC 2012 dataset [12] under the setting of 8-2 with overlapped foregrounds.

Class ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	mIoU	Imp.
Finetuning + FL	70.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	13.6	4.0	⇑ 31.7
LWF [27] + FL	**83.1**	0.2	0.0	0.0	5.0	0.0	0.0	6.8	0.0	0.0	0.0	0.0	0.0	1.1	0.0	64.5	0.3	**4.5**	**19.7**	2.8	4.2	9.2	⇑ 26.5
ILT [36] + FL	82.7	8.3	0.3	0.0	11.5	0.0	1.0	0.3	0.0	0.0	0.0	0.0	0.0	8.9	**0.5**	65.0	0.8	**15.9**	12.1	4.6	4.9	10.3	⇑ 25.4
MiB [1] + FL	82.2	26.1	**32.0**	1.7	5.8	35.3	7.8	**72.5**	58.4	4.1	0.0	0.0	15.8	14.4	**12.3**	**74.7**	**2.9**	0.0	17.5	0.0	12.7	22.7	⇑ 13.0
PLOP [11] + FL	82.6	**76.2**	**33.9**	39.1	**57.3**	55.9	39.1	71.8	48.2	0.3	0.0	**0.7**	7.7	**15.6**	0.0	61.6	0.0	0.0	13.1	**9.6**	10.5	29.7	⇑ 6.0
RCIL [53] + FL	81.1	59.2	31.9	**43.0**	**60.3**	**64.3**	**63.5**	**81.5**	**74.2**	**5.6**	0.0	**0.2**	**35.1**	4.0	0.2	66.2	**2.6**	0.0	9.0	**5.8**	**19.3**	**33.7**	⇑ 2.0
FBL (Ours)	**84.2**	**80.6**	28.7	**64.8**	54.2	**62.7**	**58.3**	66.6	**72.5**	**8.4**	0.0	0.0	**37.4**	**22.1**	0.0	**69.5**	0.0	0.0	**22.3**	1.3	**15.3**	**35.7**	–

Table 4. Comparisons of mIoU (%) on ADE20k dataset [59] under the setting of 100-10 with overlapped foregrounds.

Class ID	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100	101-110	111-120	121-130	131-140	141-150	mIoU	Imp.
Finetuning + FL	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.3	0.7	⇑ 27.2
LWF [27] + FL	0.8	0.0	0.0	0.0	0.0	0.5	1.3	0.1	0.0	0.1	0.0	0.1	0.0	1.6	9.4	0.9	⇑ 27.0
ILT [36] + FL	0.8	0.0	0.0	0.3	0.0	0.5	1.9	0.2	0.1	0.3	0.0	0.1	0.0	0.7	4.3	0.6	⇑ 27.3
MiB [1] + FL	59.7	41.3	42.4	32.4	27.9	36.4	28.7	28.9	30.1	14.7	**2.8**	**5.2**	5.8	**6.0**	**15.4**	25.4	⇑ 2.5
PLOP [11] + FL	60.7	43.4	43.8	33.7	28.8	**37.0**	**31.1**	30.3	**32.1**	15.9	2.4	**5.2**	**6.8**	2.5	**11.5**	26.1	⇑ 1.8
RCIL [53] + FL	**63.0**	**46.6**	**47.2**	**35.1**	**31.2**	36.0	30.8	**32.7**	28.1	**16.2**	0.3	**12.3**	**6.3**	2.6	3.5	**26.4**	⇑ 1.5
FBL (Ours)	**67.5**	**47.9**	**48.9**	**38.8**	**33.2**	**42.6**	**35.4**	**35.2**	**32.8**	**17.9**	**4.3**	3.1	3.3	**2.7**	0.2	**27.9**	–

essential to tackle inter-client heterogeneous forgetting via considering Non-IID distributions across local clients.

4.5. Optimization Procedure

At the beginning of each global round in each incremental task, all local clients employ Eq. (11) to calculate the average entropy of local data, and then some of local clients are randomly selected by global server $\mathcal{S}_g$ to conduct local training at each round. After these chosen clients utilize task transition monitor to accurately recognize new classes, they automatically store the global model learned at the last global round as the old model Θ^{t-1} to generate confident pseudo labels for old classes via Eq. (2), and optimize local model $\Theta_l^{r,t}$ via $\mathcal{L}_{\mathrm{obj}}$ in Eq. (10). Finally, the updated local models $\Theta_l^{r,t}$ of selected local clients are aggregated as $\Theta^{r+1,t}$ by $\mathcal{S}_g$ for the next round training. The supplementary material provides optimization procedure of our FBL model.

5. Experiments

5.1. Implementation Details

We utilize two benchmark datasets: Pascal-VOC 2012 [12] and ADE20k [59] under various experimental settings to analyze effectiveness of our FBL model. For fair comparisons with baseline ISS methods [1, 11, 27, 36, 53] under the FISS settings, we follow them to set exactly the same incremental tasks and class order, while using the identical segmentation backbone (*i.e.*, Deeplab-v3 [4] with ResNet-101 [17] pretrained on ImageNet dataset [7]). As claimed in [1, 11, 53], background pixels in the current task may belong to old classes or new classes from future tasks (*i.e.*, background has some overlap with new foreground classes in the future tasks). In the FISS, we consider more challenging settings by assigning more incremental segmentation tasks with overlapped foregrounds. On Pascal-VOC 2012 [12], 15-1, 4-4, and 8-2 settings with overlapped foregrounds respectively consist in 15 classes followed by 1 classes 5 times ($T = 6$), learning 4 classes followed by 4 classes 4 times ($T = 5$), and 8 classes followed by 2 classes 6 times ($T = 7$). Likewise, on ADE20k [59], 100-10 setting with overlapped foregrounds means 100 classes followed by 10 classes 5 times ($T = 6$).

We employ SGD optimizer with initial learning rate as 1.0×10^{-2} to train the first base task and 1.0×10^{-3} to learn incremental tasks. Considering the limitation of GPU overhead, we set initial local clients as 10, and add 4 new local clients for each task. We choose 4 local clients randomly to perform local training with 6 epochs for VOC [12] and 12 epoches for ADE20k [59]. On VOC dataset [12], we randomly select 40% images for each client in each segmentation task under 15-1 setting; otherwise, we randomly sample 50% classes from current label space $\mathcal{Y}^t$, and assign 60% samples from these classes to selected local clients under the 4-4 and 8-2 settings. For the 100-10 setting in ADE20k [59], we randomly choose 70% classes from $\mathcal{Y}^t$, and distribute them to selected clients. Following ISS methods [1, 11, 27, 36, 53], we employ mean Intersection over Union (mIoU) as metric, and evaluate mIoU of all classes after learning the last segmentation task (*i.e.*, $t = T$). This metric evaluates the effectiveness to address heterogeneous forgetting and the ability to segment new classes continually.

5.2. Comparison Performance

Experiments on Pascal-VOC 2012 [12] and ADE20k [59] are introduced to analyze superiority of our model under various settings of FISS, as shown in Tables 1∼4. Our model achieves large improvements over existing ISS methods [1, 11, 27, 36, 53] about $1.5\% \sim 51.4\%$ mIoU under various FISS settings. It illustrates the effectiveness of our model against other ISS methods to learn a global continual

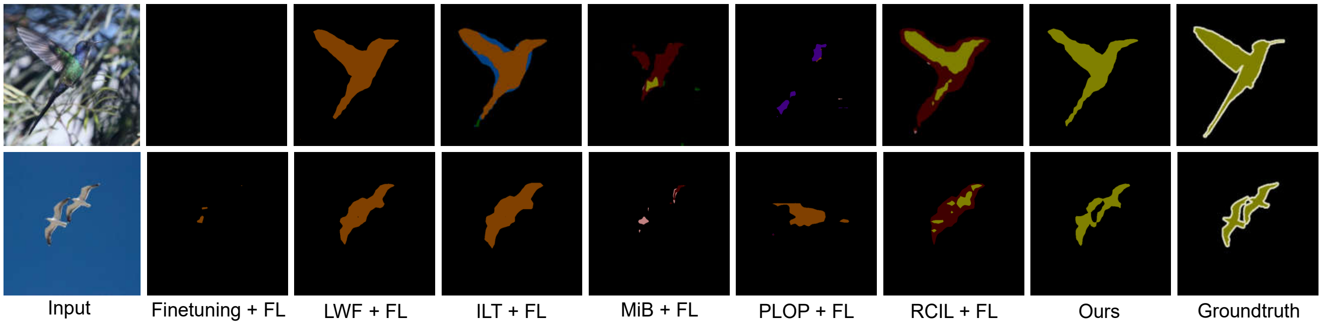

Figure 3. Visualization of some qualitative comparison results on Pascal-VOC 2012 [12] under the overlapped 4-4 setting in the FISS.

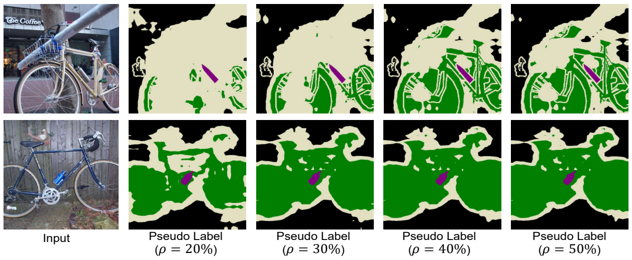

Figure 4. Visualization of some pseudo labels on Pascal-VOC 2012 [12] under the 4-4 setting with overlapped foregrounds.

Table 5. Ablation studies on Pascal-VOC 2012 [12] under the FISS.

	Variants			VOC 4-4 [12]				VOC 8-2 [12]			
Settings	APL	FSC	FRC	0-16	17-20	mIoU	Imp.	0-18	19-20	mIoU	Imp.
Our-w/oAPL	✗	✓	✓	41.3	34.3	40.0	⇑ 3.9	28.4	**10.4**	26.7	⇑ 9.0
Our-w/oFSC	✓	✗	✓	41.4	33.8	40.0	⇑ 3.9	31.6	6.8	29.2	⇑ 6.5
Our-w/oFRC	✓	✓	✗	32.6	30.8	32.3	⇑11.6	30.8	8.1	28.7	⇑ 7.0
FBL (Ours)	✓	✓	✓	**45.8**	**35.8**	**43.9**	–	**38.5**	8.3	**35.7**	–

Table 6. Task-wise comparisons of mIoU (%) on Pascal-VOC 2012 dataset [12] under the setting of overlapped 4-4 ($T = 5$).

Task ID	t=1 (Base)	t=2	t=3	t=4	t=5
Finetuning + FL	70.4	43.1	21.3	19.0	9.1
LWF [27] + FL	70.4	59.8	38.7	39.1	23.8
ILT [36] + FL	70.4	56.4	36.9	35.3	22.7
MiB [1] + FL	70.4	**64.8**	**52.8**	**47.2**	**33.0**
PLOP [11] + FL	70.4	54.2	38.3	29.4	28.1
RCIL [53] + FL	70.5	60.3	40.1	36.8	32.4
FBL (Ours)	70.4	**66.6**	**53.6**	**49.6**	**43.9**

segmentation model via collaboratively training local models under privacy preservation. Besides, it validates superiority of the proposed loss $\mathcal{L}_{\mathrm{FS}}$ and $\mathcal{L}_{\mathrm{FR}}$ to address intra-client and inter-client forgetting heterogeneity in the FISS settings. Some visualization results on Pascal-VOC 2012 [12] under the 4-4 setting are shown in Figure 3, which verifies the effectiveness of our model to address the FISS problem.

5.3. Ablation Studies

To analyze effectiveness of each module in our model, Table 5 presents ablation experiments under various FISS settings. Ours-w/oAPL, Ours-w/oFSC and Ours-w/oFRC indicate the results of our model without adaptive class-balanced pseudo labeling (denoted as APL), forgetting-balanced semantic compensation loss $\mathcal{L}_{\mathrm{FS}}$ (denoted as FSC) and forgetting-balanced relation consistency loss $\mathcal{L}_{\mathrm{FR}}$ (denoted as FRC), where Ours-w/oAPL uses constant probability threshold for all old classes to replace adaptive class-specific entropy threshold. When compared with Ours, all ablation variants severely degrade $3.9\% \sim 11.6\%$ mIoU. It verifies importance of all modules to address the heterogeneous forgetting. The proposed APL module can effectively tackle background shift via confident pseudo labels, and some confident pseudo labels are visualized in Figure 4.

5.4. Analysis of Task-Wise Comparisons

As presented in Table 6, we introduce task-wise comparison results to analyze the effectiveness of our model to address FISS settings. Our model outperforms baseline ISS methods [1, 11, 27, 36, 53] for most task-wise comparisons under the overlapped 4-4 setting. The proposed FBL model encourages local clients to learn a global incremental segmentation model cooperatively under privacy preservation. Comparisons in Table 6 show large mIoU improvements of our model to address the FISS problem over other ISS methods. When segmenting new foreground classes consecutively, our model can effectively tackle intra-client and inter-client heterogeneous forgetting on different old classes.

6. Conclusion

In this work, we propose a Federated Incremental Semantic Segmentation (FISS) problem, and develop a novel Forgetting-Balanced Learning (FBL) model to address intra-client and inter-client heterogeneous forgetting on old classes. To tackle intra-client heterogeneous forgetting, we design a forgetting-balanced semantic compensation loss and a forgetting-balanced relation consistency loss, under the guidance of adaptive class-balanced pseudo labeling. Meanwhile, we propose a task transition monitor to address inter-client heterogeneous forgetting. It can automatically recognize new classes and store the latest old global model for distillation. Comparison results demonstrate the superiority of our model to tackle the FISS problem. In the future, we will consider using only few samples of new classes to address intra-client and inter-client forgetting.

References

[1] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *CVPR*, June 2020. 2, 3, 4, 5, 6, 7, 8

[2] Sungmin Cha, beomyoung kim, YoungJoon Yoo, and Taesup Moon. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, volume 34, pages 10919–10930. Curran Associates, Inc., 2021. 3, 4

[3] Feilong Chen, Duzhen Zhang, Minglun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. VLP: A survey on vision-language pre-training. *Int. J. Autom. Comput.*, 20(1):38–56, 2023. 3

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. 1, 7

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018. 2

[6] Yang Chen, Xiaoyan Sun, and Yaochu Jin. Communication-efficient federated deep learning with layerwise asynchronous model update and temporally weighted aggregation. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10):4229–4238, 2020. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 7

[8] Jiahua Dong, Yang Cong, Gan Sun, Zhen Fang, and Zhengming Ding. Where and how to transfer: Knowledge aggregation-induced transferability perception for unsupervised domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 1

[9] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *CVPR*, pages 4022–4031, June 2020. 2

[10] Jiahua Dong, Lixu Wang, Zhen Fang, Gan Sun, Shichao Xu, Xiao Wang, and Qi Zhu. Federated class-incremental learning. In *CVPR*, June 2022. 2, 3, 6

[11] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *CVPR*, pages 4040–4050, June 2021. 2, 3, 4, 5, 6, 7, 8

[12] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2):303–338, jun 2010. 6, 7, 8

[13] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *NeurIPS*, 33:3557–3568, 2020. 1, 2

[14] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *NeurIPS*, 2022. 6

[15] Zhen Fang, Jie Lu, Feng Liu, Junyu Xuan, and Guangquan Zhang. Open set domain adaptation: Theoretical bound and algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, 32(10):4309–4322, 2021. 2

[16] Lidia Fantauzzo, Eros Fanì, Debora Caldarola, Antonio Tavera, Fabio Cermelli, Marco Ciccone, and Barbara Caputo. Feddrive: Generalizing federated learning to semantic segmentation in autonomous driving. In *Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022. 1, 2, 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016. 7

[18] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NeurIPS Workshop*, 2015. 2, 5

[19] Qirong Ho, James Cipar, Henggang Cui, Jin Kyu Kim, Seunghak Lee, Phillip B. Gibbons, Garth A. Gibson, Gregory R. Ganger, and Eric P. Xing. More effective distributed ml via a stale synchronous parallel parameter server. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 1223–1231, 2013. 2

[20] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, pages 5132–5143. PMLR, 2020. 1

[21] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3

[22] Matthias De Lange, Xu Jia, Sarah Parisot, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. Unsupervised model personalization while preserving privacy and scalability: An open problem. In *CVPR*, pages 14463–14472, 2020. 1

[23] Guangrui Li, Guoliang Kang, Wu Liu, Yunchao Wei, and Yi Yang. Content-consistent matching for domain adaptive semantic segmentation. In *ECCV*, pages 440–456. Springer, 2020. 2

[24] Guangrui Li, Guoliang Kang, Yi Zhu, Yunchao Wei, and Yi Yang. Domain consensus clustering for universal domain adaptation. In *CVPR*, 2021. 2

[25] Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *International workshop on machine learning in medical imaging*, pages 133–141. Springer, 2019. 1, 2

[26] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *ICLR*, 2021. 2

[27] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. 3, 5, 6, 7, 8

[28] Boyi Liu, Lujia Wang, Ming Liu, and Cheng-Zhong Xu. Federated imitation learning: A novel framework for cloud robotic systems with heterogeneous sensor data. *IEEE Robotics and Automation Letters*, 5(2):3509–3516, 2020. 1, 2

[29] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021. 1, 2, 3

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1, 2

[31] David Lopez-Paz and Marc' Aurelio Ranzato. Gradient episodic memory for continual learning. In *NeurIPS*, volume 30, 2017. 3

[32] Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *CVPR*, June 2018. 3

[33] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of learning and motivation*, 24:109–165, 1989. 3

[34] H Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2, 2016. 2

[35] Manan Mehta and Chenhui Shao. Federated learning-based semantic segmentation for pixel-wise defect detection in additive manufacturing. *Journal of Manufacturing Systems*, 64:197–210, 2022. 2

[36] Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. In *ICCV Workshops*, Oct 2019. 2, 3, 5, 6, 7, 8

[37] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *CVPR*, pages 1114–1124, June 2021. 3

[38] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. In *ICLR*, 2020. 2

[39] Daiqing Qi, Handong Zhao, and Sheng Li. Better generative replay for continual federated learning. *arXiv preprint arXiv:2302.13001*, 2023. 3

[40] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, July 2017. 2, 3

[41] Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 3:3, 2018. 2

[42] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *NeurIPS*, volume 30, 2017. 3

[43] Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *ECCV*, 2022. 2

[44] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *ICLR*, 2020. 1, 2

[45] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *AAAI*, volume 35, pages 10165–10173, 2021. 5

[46] Lixu Wang, Shichao Xu, Ruiqi Xu, Xiao Wang, and Qi Zhu. Non-transferable learning: A new approach for model ownership verification and applicability authorization. In *ICLR*. 2

[47] Kun Wei, Da Chen, Yuhong Li, Xu Yang, Cheng Deng, and Dacheng Tao. Incremental embedding learning with disentangled representation translation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

[48] Kun Wei, Cheng Deng, and Xu Yang. Lifelong zero-shot learning. In *IJCAI*, pages 551–557, 7 2020. 2

[49] Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. In *ICLR*, 2021. 2

[50] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2), jan 2019. 2

[51] Lu Yu, Xialei Liu, and Joost Van de Weijer. Self-training for class-incremental semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3, 4

[52] Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *ICML*, pages 7252–7261. PMLR, 2019. 2

[53] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *CVPR*, pages 7053–7064, June 2022. 2, 3, 6, 7, 8

[54] Jie Zhang, Chen Chen, Bo Li, Lingjuan Lyu, Shuang Wu, Shouhong Ding, Chunhua Shen, and Chao Wu. Dense: Data-free one-shot federated learning. In *NeurIPS*. 6

[55] Jie Zhang, Bo Li, Chen Chen, Lingjuan Lyu, Shuang Wu, Shouhong Ding, and Chao Wu. Delving into the adversarial robustness of federated learning. *arXiv preprint arXiv:2302.09479*, 2023. 6

[56] Jie Zhang, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, Lei Zhang, and Chao Wu. Towards efficient data free blackbox adversarial attack. In *CVPR*, pages 15115–15125, 2022. 2

[57] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Causaladv: Adversarial robustness through the lens of causality. 2022. 2

[58] Yonggang Zhang, Xinmei Tian, Ya Li, Xinchao Wang, and Dacheng Tao. Principal component adversarial example. *IEEE Transactions on Image Processing*, 29:4804–4815, 2020. 2

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 5122–5130, 2017. 7